Logistic Regression

Yesoda Bhargava, MPH, University of York, UK.

Introduction

- The simple and multiple regression methods are used to model the relationship between a quantitative response variable and one or more explanatory variables.
- When studying them, we attempted to estimate a population regression equation:

  $$\mu_{y|x_1,x_2,\dots,x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q \tag{1}$$

  by fitting the model

  $$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \varepsilon \tag{2}$$

  The response Y was continuous and was assumed to follow a normal distribution. We were concerned with predicting or estimating the mean value of the response corresponding to a given set of values for the explanatory variables.
- There are many situations, however, in which the response of interest is dichotomous rather than continuous. Examples of variables that assume only two possible values are disease status (the disease is either present or absent) and survival following surgery (a patient is either alive or dead).
- In general, the value 1 is used to represent a "success", or the outcome we are most interested in, and 0 represents a "failure". The mean of the dichotomous random variable Y, designated p, is the proportion of times that it takes the value 1. Equivalently,
  p=P(Y=1) = P("success")

- Just as we estimated the mean value of the response when Y was continuous, we would like to be able to estimate the probability p associated with a dichotomous response(which, of course, is also its mean) for various values of an explanatory variable. To do this, we use a technique known as *logistic regression*.
- For example of Logistic Regression, we use the birthwt dataset from MASS package.
- Response variable is low birth weight (LOW, 1=Yes, 0=No). Explanatory variable = Mother's weight at last menstrual period (LWT), Smoking status of mother (SMOK).
- Based on birth weight and smoking status of the mother we would like to see what is the probability of the child labeled as "low birth weight".
- First let us see the scatter plot of low birth weight status (LOW) with LWT.

- Note that in Fig. 1, all points lie on one of two parallel lines, depending upon whether Y, in our case, LOW, takes the values 0 or 1. There does appear to be a tendency for mother who deliver low birth weight babies to themselves have less weight reported at last menstrual period (LMP).
- Since, the two-way scatter plot is not particularly helpful, we might instead begin to explore whether an association exists between LWT and LOW by sub-dividing the sample mothers into three categories: those weighing 120 pounds or less, those weighing between 121 and 180 pounds and those weighing 181 pounds or more. We could then estimate that an infant will be low birth weight in each of these sub-groups individually.

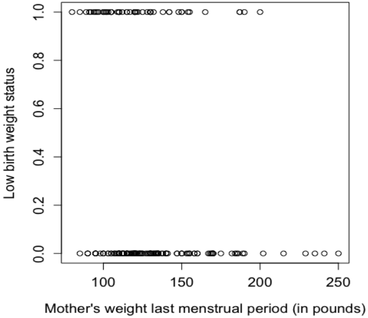


Figure 1: Scatterplot of low birth weight status versus mother's weight at last menstrual period

Table 1: Group wise probability of low birth weight

LWT (pounds)	Sample size	Number of Low birth weight	$\tilde{p}$
0-120	92	34	0.37
121-180	77	20	0.25
More than 181	16	4	0.25

We observe that the estimated probability of low-birth weight with LOWT decreases as mother's weight increases from a high of 0.37 for women weighing 120 pounds or less to a low of 0.25 for t hose weighing more than 181 pounds. Note that for the two groups (121-180) and More than 181, the proportion of LOW is the same. This is a slight discrepancy in the trend which occurs because the number of women in a particular subgroup is very small; there are only 16 women who weigh more than 181 pounds and have only 4 children who are low-birth weight. Ignoring this aberration, however, it appears that there is a relationship between these two variables. We would like to be able to use mother's LWT to help us predict the likelihood that he or she will be low birth weight.

The Logistic Function

- Our first strategy might be to fit a model of the form
  $p = \alpha + \beta x$
  where x represents the mother's weight. This is imply the standard linear regression model in which y- the outcome of a continuous, normally distributed random variable- has been replaced by p.
- As before α is the intercept of the line and β is its slope. On inspection, however, this model is not feasible. Since p is a probability, it is restricted to taking values between 0 and 1. The term $\alpha + \beta x$, in contrast, could easily yield a value that lies outside this range.
- We might try to solve this problem by instead fitting the model
  $p = e^{\alpha+\beta x}$.
  This equation guarantees that the estimate of p is positive, however, this model is also unsuitable. Although the term $e^{\alpha+\beta x}$ cannot produce a negative estimate of p, it can result in a value that is greater than 1. To accommodate this final constraint, we fit a model of the form
  $p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$
  The expression on the right, called a *logistic function*, cannot yield a value that is either negative of greater than 1; consequently, it restricts the estimated value of p to the required range.

- Recall that if an event occurs with probability p, the odds in favor of the event are p/(1-p) to 1. Thus, if a success occurs with probability

  $$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

  the odds in favor of success are

  $$\frac{p}{1-p} = \frac{e^{\alpha+\beta x}/(1+e^{\alpha+\beta x})}{1/(1+e^{\alpha+\beta x})} \qquad = e^{\alpha+\beta x}.$$

- Taking the natural logarithm of each side of this equation,

  $$\ln \frac{p}{1-p} = ln(e^{\alpha+\beta x})$$

  $$= \alpha + \beta x$$

- Thus, modeling the probability p with logistic function is equivalent to fitting a linear regression model in which the continuous response has been replaced by the logarithm of the odds of success for a dichotomous random variable.
- Instead of assuming that the relationship between p and x is linear, we assume that the relationship between ln(p/(1-p)) and x is linear. The technique of fitting a model of this form is known as logistic regression.

The Fitted Equation

- In order to use a mother's weight to help us to predict the likelihood that the new born will be low birth weight, we fit the model

  $ln\frac{p}{1-p} = \hat{\alpha} + \hat{\beta}x$
- Although we categorized LWT into three different intervals when exploring its relationship with LOW, we use the original continuous random variable for the logistic regression analysis. As with a linear regression model, $\hat{\alpha}$ and $\hat{\beta}$ are estimates of the population coefficients.
- However, we cannot apply the method of least squares which assumes that the response is continuous and normally distributed, to fit a logistic model; instead, we use maximum likelihood estimation. This technique uses the information in a sample to find the parameter estimates that are most likely to have produced the observed data.
- For the sample of 189 low birth weight infants, the estimated logistic equation is $ln\frac{\hat{p}}{1-\hat{p}} = 0.998 - 0.014 * LWT$
- The coefficient of LWT implies that for each one pound increase in mother weight at LMP, the log odds that the infant is low birth weight decrease by 0.014 on average. When the log odds decrease, the probability of p decreases as well.

- In order to test $H_0 : \beta = 0$, the null hypothesis that there is no relationship between p and x, against the alternative hypothesis $H_A : \beta \neq 0$, we need to know the standard error of the estimate $\hat{\beta}$. Then if, H_0 is true, the test statistic $z = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})}$ follows a standard normal distribution. It turns out that the coefficient of LWT is infact significantly different from 0 at the 0.05 level; thus we conclude that in the underlying population of mothers weight, the probability of low birth weight decreases as the mother's weight at LMP increases.

- In order to estimate the probability that an infant of a mother with a given LWT is low birth weight, we simply substitute the appropriate value of LWT into the preceding equation. To estimate the probability that a woman weighing 100 pounds gives birth to a low birth weight baby, for example, we substitute value LWT=100 to find $ln \frac{\hat{p}}{1-\hat{p}} = 0.998 - 0.014 * 100 = 0.998 - 1.4 = -0.402$
- Taking the antilogarithm each side of the equation $\frac{\hat{p}}{1-\hat{p}} = exp^{-0.402} = 0.668$
- Finally solving for $\hat{p}$ $\hat{p} = \frac{0.668}{1.668} = 0.395$
- The estimated probability that a woman weighing 100 pounds gives birth to a low birth weight infant is 0.395.

- If we calculated the estimated probability $\hat{p}$ for each observed value of LWT and plotted $\hat{p}$ versus LWT, the result would be the curve shown in Fig. 2.

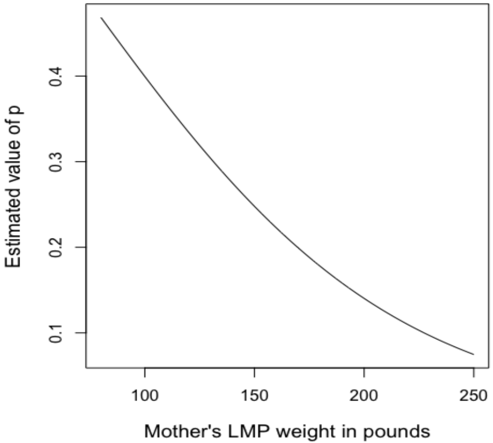


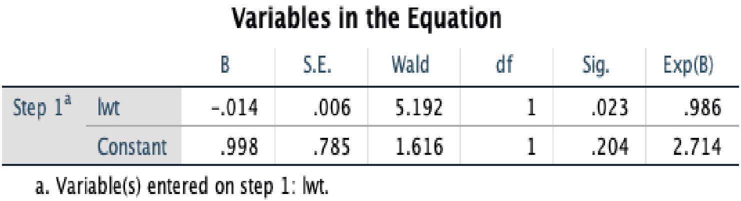
Figure 2: Scatterplot of low birth weight status versus mother's weight at last menstrual period

Inference for Logistic Regression

- Here, we will further the discussion on statistical inference for logistic regression.
- Statistical inference for logistic regression is very similar to statistical inference for simple linear regression. We calculate estimates of the model parameters and standard errors for the estimate.
- Confidence intervals are formed in the usual way, but we use standard normal z* values rather than critical values from t distribution. As seen before, the ratio of estimate to the standard error of the estimate is the basis of the hypothesis test.
- Often the test statistics are given as squares of these ratios, and in this case the P-values are obtained from the chi-square distributions with 1 degree of freedom.
- The statistic z is sometimes called the **Wald** statistic. Output from some statistical software reports the significance test results in terms of the square of the z statistic. This statistic is called the chi-square statistic. Note that in this case, the P-values are obtained from chi-square distributions with 1 degree of freedom.

Confidence intervals and significant tests for Logistic Regression Parameters

- A level C confidence interval for the slope β_1 is $b_1 \pm z * SE_{b_1}$. The ratio of the odds for a value of the explanatory variable equal to x+1 to the odds for a value of the explanatory variable equal to x is the odds ratio.
- A level C confidence interval for the odds ratio e^{β_1} is obtained by transforming the confidence interval for the slope $(e^{b_1 - z*SE_{b_1}}, e^{b_1 + z*SE_{b_1}})$
  In these expressions z* is the value for the standard normal density curve with area C between -z* and z*.
- As mentioned before to test the hypothesis $H_0 : \beta_1 = 0$, the test statistic $z = \frac{b_1}{SE_{b_1}}$ is computed.
- We have expressed the hypothesis-testing framework in terms of the slope β_1 because this form closely resembles with simple linear regression. In many applications, however, the results are expressed in terms of the odds ratio. A slope of 0 is the same as as an odds ratio of 1, so we often express the null hypothesis of interest as "the odds ratio of 1". This means that the two log odds are equal and the explanatory variables are not useful for predicting the odds.
- Fig. 3 gives the output from SPSS for the low birth weight example.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	lwt	-.014	.006	5.192	1	.023	.986
	Constant	.998	.785	1.616	1	.204	2.714

a. Variable(s) entered on step 1: lwt.

Figure 3: Logistic Regression output from SPSS for the low birth weight example.

The parameter of the estimates are given as $b_0 = 0.998$ and $b_1 = -0.014$. The standard errors are 0.785 and 0.006 respectively. A 95% confidence interval for slope is $b_1 \pm z * SE_{b_1} = -0.014 \pm 1.96 * 0.006$

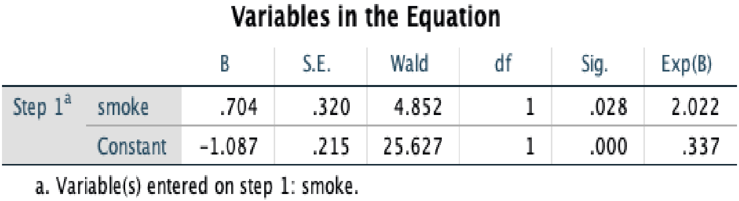
$$= -0.014 \pm 0.00176.$$

We are 95% confident that the slope is between (-0.02576, -0.00224). The output provides an odds ratio of 0.986 but does not provide the confidence interval. This is easy to compute from the interval of the slope.

$$(e^{b_1 - z * SE_{b_1}}, e^{b_1 + z * SE_{b_1}}) = (e^{-0.02576}, e^{-0.00224})$$

$$= (0.984, 0.987)$$

- For this problem we would report: "The odds for low-birth weight decrease by a factor of 0.986 for each unit increase in mother's weight at last menstrual period (X^2 = 5.192, p=0.023; 95% CI = 0.984 to 0.987).
- In applications such as these, it is standard to use 95% for the confident coefficient. With this convention, the confidence interval gives us the result of testing the null hypothesis that the odds ratio is 1 for s significance level of 0.05. If the confidence interval does not include 1 we reject the null hypothesis and conclude that the odds of low birth weight is different with increase in mother's LMP weight. If the interval does include 1, the data do not provide evidence to show that mother's weight at LMP has any association with low birth weight baby.
- Now, we regress low birth weight on smoking status of the mother and see how it works. Fig. 4 shows the SPSS output for logistic regression.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	smoke	.704	.320	4.852	1	.028	2.022
	Constant	-1.087	.215	25.627	1	.000	.337

a. Variable(s) entered on step 1: smoke.

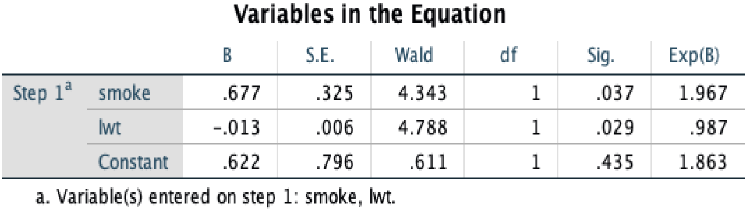
Figure 4: Logistic Regression output from SPSS for the low birth weight example using smoking status as the predictor.

For this case we can see that the variable smoking is significant with a p-value of 0.028and standard error of 0.320. We compute the 95% CI for the slope as before $0.704 \pm 1.96 \times 0.320$. This gives us (.0768, 1.33). Using these we can compute the 95% CI for the odds,

$(e^{0.0768}, e^{1.33}) = (1.079, 3.78)$. For this case we would report, ”Women who smoke are more likely to have low birth weight babies than women who do not smoke (odds ratio 2.022, 95% CI = 1.08to 3.78).

Multiple Regression Analysis

- We have seen individually how mother's LMP weight and smoking status affect the risk of low birth weight in infants, now we wish to see how together they affect the outcome. We use **Multiple Logistic Regression** to answer this question.
- Generating computer output is easy, the statistical concepts are similar. The results are shown in Fig. 4.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	smoke	.677	.325	4.343	1	.037	1.967
	lwt	-.013	.006	4.788	1	.029	.987
	Constant	.622	.796	.611	1	.435	1.863

a. Variable(s) entered on step 1: smoke, lwt.

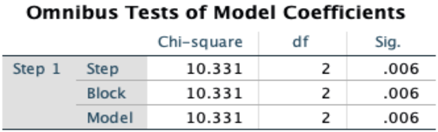
Figure 5: Logistic Regression output from SPSS for the low birth weight example using smoking status and LWT as the predictor.

The fitted model is log(ODDS)= $b_0 + b_1 LWT + b_2 SMOK$

$$= 0.622 \text{ -}0.013 \text{ LWT} + 0.677 \text{ SMOK}.$$

When analysing data using multiple regression we examine the hypothesis that all of the regression coefficients for the explanatory variables are zero. We do the same for logistic regression.

- The hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is tested by using a chi-square statistic with 2 degrees of freedom. Fig. 5 shows this result. This is called as the Omnibus Tests of model coefficients.



Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10.331	2	.006
	Block	10.331	2	.006
	Model	10.331	2	.006

Figure 6: Omnibus Tests of model coefficients.

- In this model we have added the two predictor variables at same time so there is only one step and one block hence, the Model chi-square and df are the same. In this case, we can see that the chi-square is 0.006 and the P-value is 0.006. We reject H_0 and conclude that one or more of the explanatory variables can be used to predict the odds of low birth weight.
- We now examine the coefficients for each variable and the test that each of these is zero. The P-values are 0.037 and 0.029. so, we can reject $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$

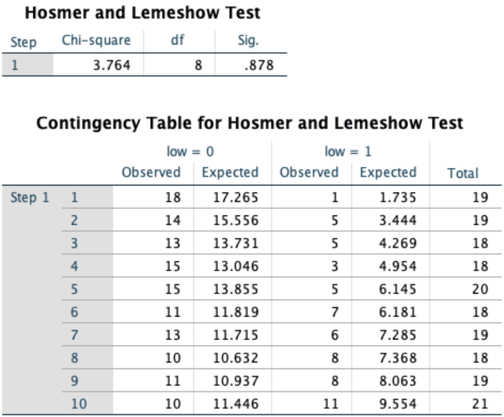
The Hosmer-Lemeshow goodness of fit test for logistic regression

- The Hosmer-Lemeshow test is used to determine the goodness of fit of the logistic regression model. Essentially it is a chi-square goodness of fit test for grouped data, usually where the data is divided into 10 equal subgroups.
- The observed and expected number of cases in each group is calculated and a Chi-squared statistic is calculated as follows:

  $$\chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1 - E_g/n_g)}$$

  where O_g signifies the observed events, E_g signifies the expected events and n_g signifies number of observations for the gth group and G is the total number of groups. The test statistic follows a chi-squared distribution with G-2 degrees of freedom.
- A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.
- The Hosmer-Lemeshow test needs to be used with caution. It tends to be highly dependent on the groupings chosen, i.e. one selection of groups can give a negative result while another will give a positive result. Also when there are too few groups (5 or less) then usually the test will show a model fit.
- Remember,for HL test the H_0 : No lack of fitness of model/Model is a good fit. H_A: Lack of fitness of model/Model is not a good fit.

For our example, these are the results of the HL test.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.764	8	.878

Contingency Table for Hosmer and Lemeshow Test

		low = 0		low = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	18	17.265	1	1.735	19
	2	14	15.556	5	3.444	19
	3	13	13.731	5	4.269	18
	4	15	13.046	3	4.954	18
	5	15	13.855	5	6.145	20
	6	11	11.819	7	6.181	18
	7	13	11.715	6	7.285	19
	8	10	10.632	8	7.368	18
	9	11	10.937	8	8.063	19
	10	10	11.446	11	9.554	21

Figure 7: Results of Hosmer-Lemeshow goodness of fit test for logistic regression for low birth weight example.

We see that the p-value for the test is 0.878 indicating that we cannot reject the null hypothesis. Hence, we say that the model provides a good fit to the data.

- The technique discussed above is not free from its criticisms. Many statisticians have found it to be unsatisfactory for many reasons.
- The most troubling problem is that results can depend markedly on the number of groups, and there's no theory to guide the choice of that number. This problem did not become apparent until software packages started allowing you to specify the number of groups, rather than just using 10.
- If the HL test is no good, then how can we assess the fit of the model? It turns out that there's been quite a bit of recent work on this topic.
- The alternatives to the HL test generally fall into two categories: tests that do not require grouping of the data, and tests that propose methods of grouping that are different than HL.
- There are four ungrouped goodness of fit tests: standardized Pearson, unweighted sum of squared residuals, Stukel's test, and the information matrix test. We try to understand the standardized Pearson test. It's familiar, relatively easy to compute, and has pretty good performance under a range of conditions.

ROC curve

- One way to measures of fit for the logistic regression model is the ROC curve.
- The receiving operating characteristic is a measure of classifier performance.
- Using the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive, we generate a graphic that shows the trade off between the rate at which you can correctly predict something with the rate of incorrectly predicting something.
- Ultimately, we're concerned about the area under the ROC curve, or AUROC. That metric ranges from 0.50 to 1.00

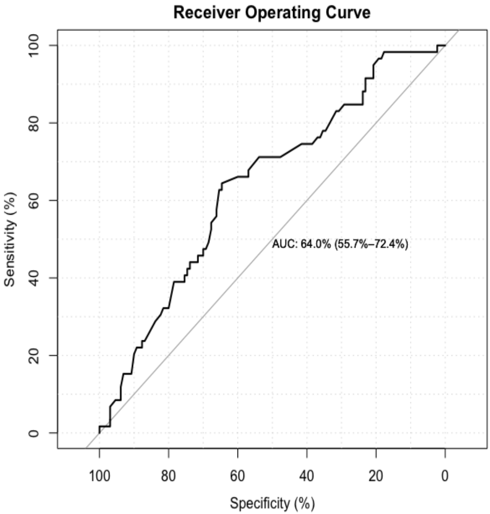


Figure 8: ROC curve for logistic regression model.

Pseudo R^2

- Unlike linear regression with ordinary least squares estimation, there is no R2 statistic which explains the proportion of variance in the dependent variable that is explained by the predictors. However, there are a number of pseudo R2 metrics that could be of value. Most notable is McFadden's R2.
- The measure ranges from 0 to just under 1, with values closer to zero indicating that the model has no predictive power.
- The value is 0.04402462 indicating a poor fit.

Sources referenced and recommended reading

- Evaluating Logistic Regression Models
- The Hosmer-Lemeshow goodness of fit test for logistic regression
- Why I Don't Trust the Hosmer-Lemeshow Test for Logistic Regression
- 1Goodness of Fit in Logistic Regression
- Hosmer-Lemeshow Test
- Hosmer-Lemeshow Goodness-of-Fit Test
- Inputs takes from Principles of Biostatistics, Marcello Pagano, Kimberlee Gauvreau.
- Introduction to the practice of statistics, David S. Moore, Groege P. McCabe.